

DOCUMENT RESUME

ED 333 046

TM 016 615

AUTHOR Linacre, John Michael
TITLE Structured Rating Scales.
PUB DATE Apr 91
NOTE 15p.; Paper presented at the International Objective Measurement Workshop (6th, Chicago, IL, April 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Equations (Mathematics); *Mathematical Models; *Rating Scales; Scaling; Test Construction; *Test Format; Test Items

ABSTRACT

A rating scale can be expressed as a chain of dichotomous items. The relationship between the dichotomies depends on the manner in which the rating scale is presented to the test taker. Three models for ordered scales are discussed. In the success model, which represents growth, the lowest or easiest category is presented first. If the test taker succeeds, consecutively higher, more difficult categories are presented until the test taker finally fails or all categories have been presented. In the failure model, which represents mastery, the highest and most difficult category is presented first. If the test taker fails, consecutively lower and easier categories are presented until the test taker finally succeeds or all categories have been presented. In the D. Andrich model, which represents attitude and other closed rating scales, all categories are presented simultaneously and the test taker selects the one most germane. One table and four graphs are included. (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

JOHN M. LINACRE

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Structured Rating Scales

by

John Michael Linacre

MESA Psychometric Laboratory
Department of Education
University of Chicago

Paper presented at
Sixth International Objective Measurement Workshop
Chicago, Illinois
April 1991

BEST COPY AVAILABLE

ED333046

T11016615

Abstract

A rating scale can be expressed as a chain of dichotomous items. The relationship between the dichotomies depends on the manner in which the rating scale is presented to the test-taker. Three models for ordered scales are discussed. In the Success model, the lowest or easiest category is presented first. If the test-taker succeeds, consecutively higher, more difficult categories are presented until the test-taker finally fails or all categories have been presented. In the Failure model, the highest and most difficult category is presented first. If the test-taker fails, consecutively lower and easier categories are presented until the test-taker finally succeeds or all categories have been presented. In the Andrich model, all categories are presented simultaneously and the test-taker selects the one most germane.

Key words: Rasch Measurement, Rating Scales, Partial Credit, Guttman Scales, Mastery, Growth

Introduction

Representing successively better levels of performance on an item by means of a hierarchical scale enables more information to be obtained from an item than would be provided by any single "more/less" dichotomy. Such polytomous items are often used for attitude scales. It may also happen, however, that a cluster of interdependent dichotomous items, or a sequence of items in an adaptive test, have a relationship so tightly hierarchical that they can represent levels of performance on one super-item. Modelling interdependent items as defining steps along the rating scale of one super-item allows the analyst the opportunity to eliminate the effects of local interdependencies among related items.

When a hierarchical polytomous item is constructed, clear thinking is needed to specify the relationship between its levels. In choosing between the models presented here, does success at a given level imply that success has **necessarily** been achieved at all lower levels, or does failure at a given level imply that all higher levels have **necessarily** been failed, or does the set of levels represent a **complete** set of levels of success for the item? Would the addition of extra levels at either extreme of the rating scale alter the relationship between existing levels?

These considerations lead to three models for how the performance levels underlying a rating scale can be defined in terms of co-dependent dichotomous steps. In the examples, each model implements three sequential dichotomous steps, with step 1 being the lowest or easiest one. It is the location of "M", representing steps not attempted, which is decisive in differentiating between the models.

1. Glas-Verhelst "Success" Model for Growth

Super-Item	Dichotomous Items			Interpretation	
Rating Level	Step 1	Step 2	Step 3	Item-Steps Tried	Successes
3	1	1	1	3	3
2	1	1	0	3	2
1	1	0	M	2	1
0	0	M	M	1	0
Start ↑	Start ⇒				

Legend: 1=Passed, 0=Failed, M=Not encountered, i.e. Missing.

The "Success" model (Glas and Verhelst, 1989) represents **growth**. Higher steps can only be attempted after success has been achieved on lower ones. Persons who fail ("0") on Step 1, do not qualify for Step 2 or Step 3, and so have no opportunity to respond to them. This results in missing ("M") responses, rather than pre-determined failures. Each rating level implies a particular count of dichotomous item-step successes based on the administration of a particular number of steps. Introducing additional, **higher**, levels does not alter the relationship between the existing lower levels. For given person measures, the estimation of an item-step calibration does not alter the calibration any lower step.

The local independence, necessary for a Rasch model (Rasch 1960, 1980) is present, because, for those attempting any level, whether that next steps is a success or failure is not pre-determined by the previous success on lower levels. Each next step calibration can take any value, and is estimated as though the steps are independent items. There is only procedural dependency, namely that solely those who succeed at all lower steps have the opportunity to try the next step. Once the decision has been made to administer the next step, the success obtained on the previous levels becomes irrelevant, and the expected next response depends only on the ability of the person and the difficulty of that next step.

The algebraic representation of the probability of success on a step remains based on the ratio between successes and failures for all who encountered this step, just as it is for independent dichotomous items. For those, and only those, who succeed in reaching step j ,

$$\log\left(\frac{P_{nij}}{1-P_{nij}}\right) = B_n - D_{ij} \quad j=1, \text{Min}(X_{ni}+1, M_i) \quad (1)$$

where P_{nij} is the probability of success of person n , with ability measure B_n on the j th step with difficulty calibration D_{ij} of hierarchical item i , which has M_i steps. X_{ni} is the observed rating of person n on item i .

This is realized analytically in two ways:

a) For prediction, the expected item score is the sum of the levels (i.e. steps) multiplied by their probabilities of being observed. The probability of a level is obtained by multiplying together the probabilities that someone succeeds on all lower steps as well as this one, but also fails on the next higher step:

$$P_{ni} = \sum_{j=1}^{M_i} \left[j (1 - P_{nij+1}) \prod_{k=1}^j P_{nik} \right] \quad (2)$$

where P_{ni} is the expected score by person n on item i ,
and the unobservable step denoted by M_i+1 has zero probability.

Figure 1 shows the measure-to-expected raw score ogive for a sample data set calibrated according to this model.

b) For maximum likelihood estimation, the structure of the observed data must be maintained. The item score level observed specifies the number of dichotomous item steps administered. We do not have the option to alter the number of dichotomous steps (i.e. change the length of the test) during estimation, but must perform estimation strictly on the number of independent item-steps administered. Maximum likelihood estimation is based on the expected score for the observed item-steps:

$$P_{ni} = \sum_{j=1}^{\text{Min}(X_{ni}+1, M_i)} P_{nij} \quad (3)$$

2. Linacre "Failure" Model for Mastery

Super-Item	Dichotomous Items			Interpretation	
Rating Level	Step 1	Step 2	Step 3	Item-Steps Tried	Successes
Start ↓	≡ Start				
3	M	M	1	1	1
2	M	1	0	2	1
1	1	0	0	3	1
0	0	0	0	3	0

The "Failure" model represents **mastery**, as in answer-until-correct and other situations where simpler tasks are presented only when complex tasks have been failed (Linacre 1990). The "Failure" model is the reverse of the "Success" model. Lower dichotomous steps are only administered when success has **not** been achieved on higher steps. Step 2 is administered only when Step 3 has been failed. Thus a success on Step 3 means that Steps 1 and 2 are not administered, and are thus missing, "M", rather than pre-determined successes. In this case, **lower** levels can be added without altering the structure of the upper scale. This, too, is a Rasch formulation, because of the independence of success on each lower step. The algebraic representation of the probability of success on a step is identical to the Success model. For those and only those who fail down to step j ,

$$\log\left(\frac{P_{nij}}{1-P_{nij}}\right) = B_n - D_{ij} \quad j = \text{Max}(1, X_{ni}), M_i \quad (4)$$

with the same notation as (1).

But this model is realized analytically in ways which **differ** from the "Success" model:

a) For prediction, the expected item score is the sum of the levels multiplied by their probabilities of being observed. The probability of a level is obtained by multiplying together the probabilities that someone **fails on all higher steps**, but succeeds on this one:

$$P_{ni} = \sum_{j=1}^{M_i} \left[j P_{nij} \prod_{k=j+1}^{M_i+1} (1 - P_{nik}) \right] \quad (5)$$

Figure 2 shows the measure-to-expected raw score ogive for the same sample data set as used in Figure 1, but calibrated according to this model.

b) For maximum likelihood estimation, the considerations of the "Success" model apply, but modelled as failures instead of successes:

$$P_{ni} = \sum_{j=\text{Max}(1, x_{ni})}^{M_i} P_{nij} \quad (6)$$

3. Andrich "Rating Scale" and Masters "Partial Credit" Models

Super-Item	Dichotomous Items			Interpretation	
Rating Level	Step 1	Step 2	Step 3	Item-Steps Tried	Successes
3	1	1	1	3	3
2	1	1	0	3	2
1	1	0	0	3	1
0	0	0	0	3	0

These models represent attitude and other **closed rating scales**, where the selection of a level implies a position along a completely defined continuum (Andrich 1978, Masters 1982, Wright & Masters 1982). The levels specify a conceptual hierarchy, but do not imply a path the person **must** follow in order to reach a particular level. The scale can be understood as ascent from the bottom, descent from the top, or digression from the center. The usual algebraic expression as ascent from the bottom is mathematical convenience.

This is a Rasch model because the parameter corresponding to any step is separable from those of the other steps. One can estimate the step difficulties in ascending order, beginning with the observed frequencies of the bottom two categories, and proceed up the steps without knowledge of the frequencies of observations in the higher categories. Or one can do the reverse. The parameters for other steps can be conditioned out of the estimation equation for any particular step. The pairwise approach to estimation (Wright & Masters, p.69-72) takes advantage of this.

Making a posterior change in the structure of this scale, however, alters the statistical relationships between all levels of the scale. Such a change may become necessary when inspection of the data discloses that some rating levels have rarely or never been observed.

The algebraic representation of the probability of success on a step is:

$$\log \left(\frac{P_{nij}}{P_{nij-1}} \right) = B_n - D_{ij} \quad j=1, M_i \quad (7)$$

where P_{nij} is the probability of the observation of a rating of j for person n on item i . P_{ni0} is the probability of a rating of 0.

Since all levels participate in every observation, the formulation for the expected score and for determining the maximum likelihood of the data is

$$P_{ni} = \sum_{j=1}^{M_i} j P_{nij} \quad (8)$$

Figure 3 shows the measure-to-expected raw score ogive for the same sample data set as used in Figure 1, but calibrated according to this model.

Implications for calibration

Numerically equivalent observations imply significantly different calibrations for these three Rasch scale models. Table 1 contains the calibrations. Figure 4 presents these calibrations graphically.

Excursus on Guttman "Graded Response" Model (Samejima/Thurstone/McCullagh)

Super-Item	Dichotomous Items			Interpretation	
Rating Level	Step 1	Step 2	Step 3	Item-Steps Tried	Successes
3	1	1	1	3	3
2	1	1	0	3	2
1	1	0	0	3	1
0	0	0	0	3	0

In this approach, success at a particular level implies success at **all** levels below and also failure at **all** levels above (Thurstone 1928, Guttman 1950, Samejima 1972, McCullagh 1980). The "graded response" model of the difficulty of a step for the Guttman model is:

$$\log \left(\frac{\sum_{k=j}^{M_i} P_{nik}}{\sum_{k=0}^{j-1} P_{nik}} \right) = B_n - D_{ij} \quad j=1, M_i \quad (9)$$

Lack of the local independence among the steps necessary for Rasch measurement occurs because the frequencies of all categories must be known before an estimate can be made of even one step. The Andrich model retains the Guttman form, but with local independence of the steps.

The area in which proponents of this Guttman model claim strength is in the analysis of scales with many conceptual categories. Such a scale could be presented as a line along which a test-taker makes a mark. But such scales can be successfully analyzed as Andrich scales in which the relationship between the position on the line and the challenge presented is either explicitly hypothesized or empirically modelled from the data (Linacre 1988).

References:

- Andrich D (1978) A rating formulation for ordered response categories. *Psychometrika* 43(4): 69-81.
- Glas C A W, Verhelst N D (1989) Using the Rasch model for dichotomous data for analyzing polytomous responses. Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, California.
- Linacre J M (1988) Rating scales with many categories. Paper presented at the Mid-West Objective Measurement Seminar, Chicago, Il. May.
- Linacre J M (1990) Modelling rating scales. Paper presented at AERA Annual Meeting, Boston, 1990.
- Guttman L (1950) The basis for scalogram analysis. In *Measurement and Prediction*, volume IV of *The American soldier*, Ed. S. Stouffer et al. New York: John Wiley.
- Masters G N (1982) A Rasch model for partial credit scoring. *Psychometrika* 47:149-174.
- McCullagh P (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society B* 42(2): 109-142.
- Rasch G (1960, 1980) Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.
- Samejima F (1972) A general model for free-response data. *Psychometrika* 37(2) monograph supplement No. 18.
- Thurstone L L (1928) Attitudes can be measured. *American Journal of Sociology*. 33: 529-554.
- Wright B D (1991) Private communication with author.
- Wright B D, Linacre J M (1991) BIGSTEPS Rasch analysis computer program. Chicago: MESA Press
- Wright B D, Masters G N (1982) *Rating Scale Analysis*. Chicago: MESA Press.

Easy Item							
Level	Frequency	Item-Step Calibrations					
		Success		Failure		Andrich	
2	10	10/10	0.0	10/11	-0.1	10/10	0.0
1	10	20/1	3.0	10/1	2.3	10/1	2.3
0	1	-	-	-	-	-	-

Medium Item							
Level	Frequency	Item-Step Calibrations					
		Success		Failure		Andrich	
2	10	10/1	2.3	10/11	-0.1	10/1	2.3
1	1	11/10	0.1	1/10	-2.3	1/10	-2.3
0	10	-	-	-	-	-	-

Hard Item							
Level	Frequency	Item-Step Calibrations					
		Success		Failure		Andrich	
2	1	1/10	-2.3	1/20	-3.0	1/10	-2.3
1	10	11/10	0.1	10/10	0.0	10/10	0.0
0	10	-	-	-	-	-	-

Table 1. Scale calibrations for representative items, according to the three models, when administered to 21 persons of measure 0. The calibrations are shown on a ratio scale (in fractional odds units) and on an interval scale (in decimal logits).

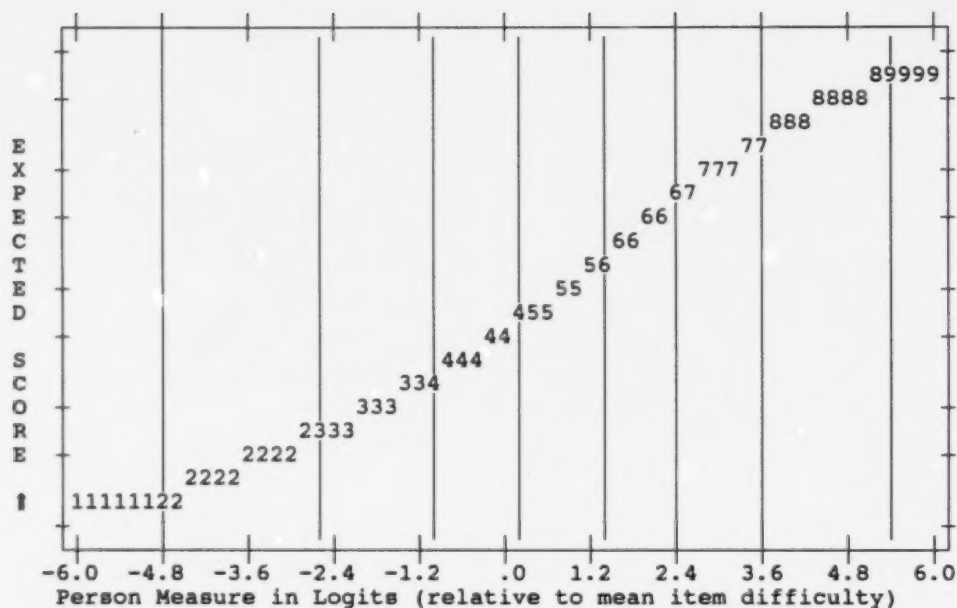


Figure 1. Expected raw score on a nine category scale, calibrated according to the Success model, using the BIGSTEPS computer program. Digits printed within plot indicate the expected score rounded to the nearest integer.

Note the somewhat concave form of the ogive for this data, with category 6 at the steepest part of the ogive.

Data supplied by H. I. Braun contributed to this figure.

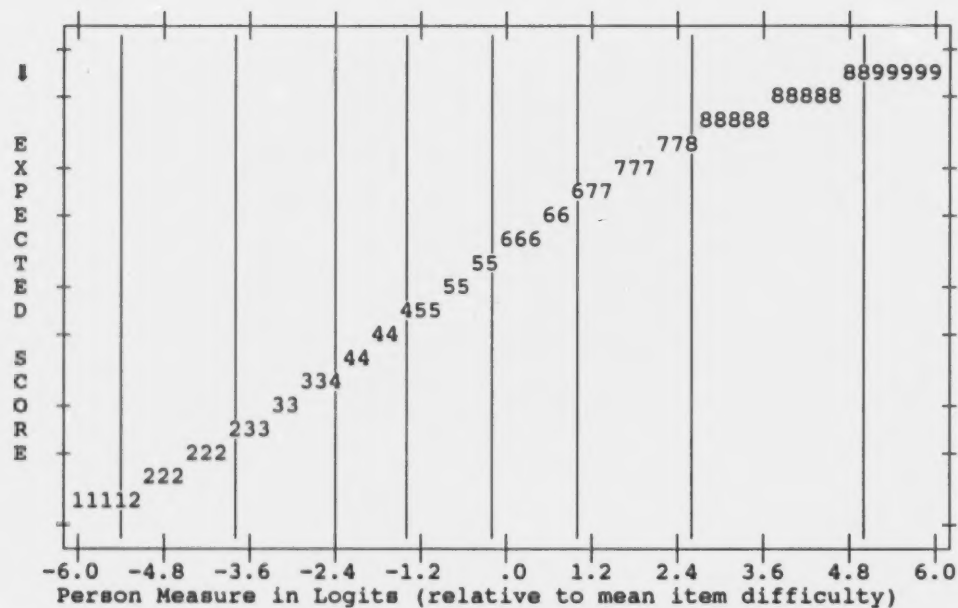


Figure 2. Expected raw score on a nine category scale, calibrated according to the **Failure** model, using the BIGSTEPS computer program. Note the somewhat convex form of the ogive for this data, with category 4 at the steepest part of the ogive. Data supplied by H. I. Braun contributed to this figure.

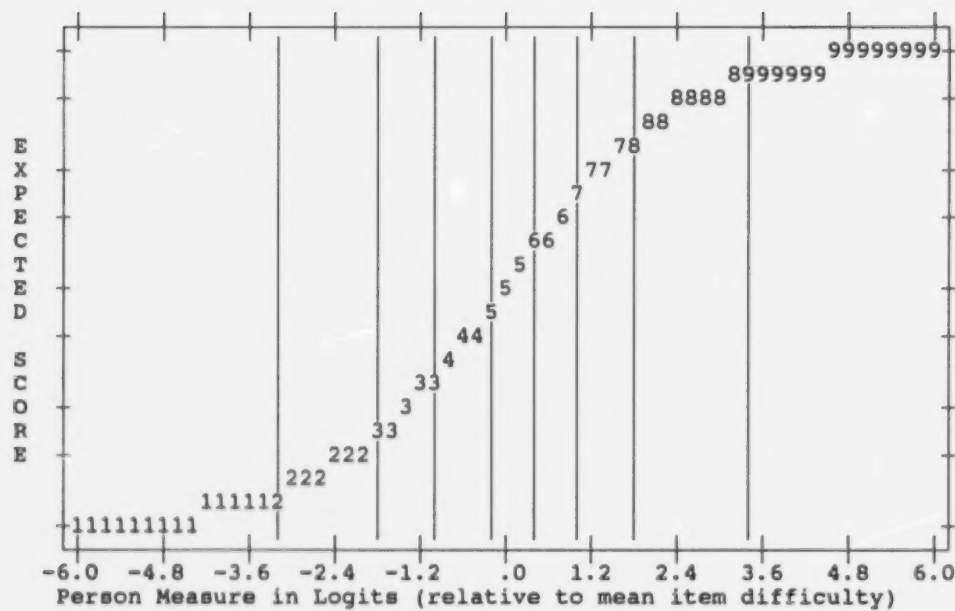


Figure 3. Expected raw score on a nine category scale, calibrated according to the **Andrich** model, using the **BIGSTEPS** (1991) computer program. Note the somewhat symmetrical form of the ogive for this data, with category 5 at the steepest part of the ogive. Data supplied by H. I. Braun contributed to this figure.

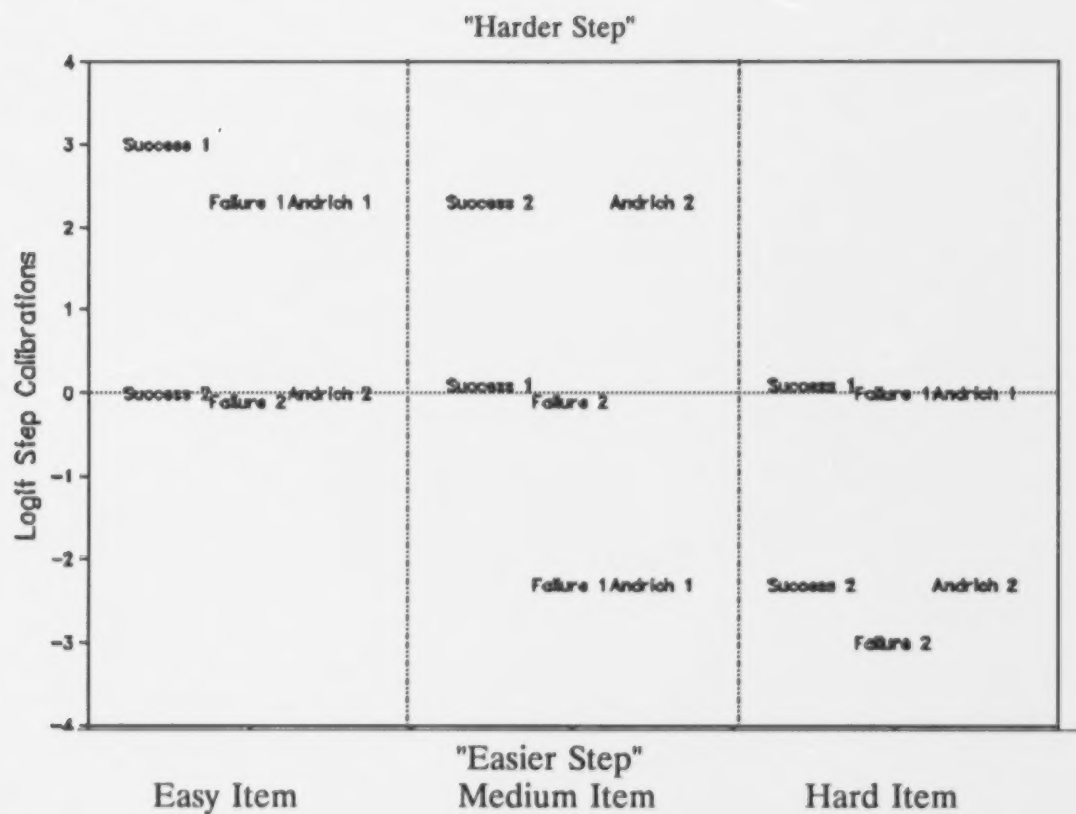


Figure 4. Plot of logit calibrations for easy, medium and hard items, with 2 steps (3 categories), according to the three models. The step calibrations are "disordered" for these easy and hard items.